



Cnli0030428.doc  
January 18, 2002

2153  
#3  
**RECEIVED**

**JAN 21 2003**

Confirmatory License

Technology Center 2100

Application For: Surveillance Monitoring and Automated Reporting Method for Detecting Data Changes

Inventor(s): Andrew H. Quintero, Jeffrey S. Fedor, Alan G. Quan, Karen Richardson, Donald W. Scott, Ken A. Piper

Serial No.: 09/824,340

Contract No.: F04701-00-C-0009

Filing Date: 04/02/01

Contractor: The Aerospace Corporation

The invention identified above is a "Subject Invention" under Patent Rights clause FAR 52.227, which is included in Contract No. F04701-00-C-0009, with the Department of the Air Force. This document confirms the rights acquired by the Government through the referenced clause, including a paid-up license in this invention, corresponding patent applications and in any resulting patents.

The Government is hereby granted an irrevocable power to inspect and make copies of the above-identified patent application.

Signed this 23 day of January 2002

Contractor: The Aerospace Corporation

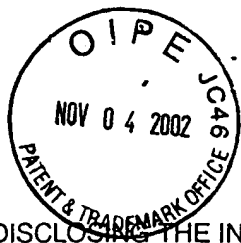
By:

Gordon J. Louttit

Senior Vice President, General Counsel and Secretary

Business Address: P.O. Box 92957, M1/040  
Los Angeles, CA 90009

*An Equal Opportunity Employer*



## INVENTION DISCLOSURE

**RECEIVED**

**JAN 21 2003**

**Technology Center 2100**

### IN DISCLOSING THE INVENTION

1. Title: Provide a brief title for the invention.
2. Technology: Indicate which Patent Review Committee should review this disclosure:  
☐ Electronics    ☒ Systems    ☐ Physics    ☐ Applied Mechanics
3. Abstract: Write a short paragraph discussing the invention, its components, and advantages.
4. Background: If you know of any highly relevant prior art, discuss the prior art's problems and methods without referring to the invention. Attach copies.
5. Invention Summary: Describe basic components of the invention and how the invention solves those problems.
6. Invention Description: Describe in detail the invention and its components while referring to elements shown in attached drawing(s).
7. Prior Disclosures: List and attach copies of all disclosures, public or private, of the invention made outside of The Aerospace Corporation.
8. Exploitation: List features of the invention that would give it a competitive advantage in the marketplace, if any, and discuss its potential uses and specific markets.
9. Funding: Was this invention government funded?
10. Execution: Sign the invention disclosure and have it witnessed by others who understand the invention. Send the invention disclosure to the Office of the General Counsel, Attn: Technical Counsel, M1/040.

**Title:** Method and Apparatus for Monitoring Changes in Information in Data Processing Systems and for Providing Scheduled Change Detection Notification and Displayed Results. (Surveillance, Monitoring, and Automated Reporting Tool for Changes in Observable Websites [SmartCow]).

**Technology:** Aerospace - developed Software executed on two computing systems, using world wide web server, relational database, and Java technologies.

**Abstract:** In this invention disclosure we describe a Method and Apparatus for Monitoring Changes in Information in Data Processing Systems and for Providing Scheduled Change Detection Notification and Displayed Results. (Surveillance, Monitoring, and Automated Reporting Tool for Changes in Observable Websites [SmartCow]).

**Background:** This project was an outgrowth of the Computer Systems Division's IR&D projects in intelligent software agent and information system technologies and requested by Andrew Quintero, to provide the functions described below.

### **Invention Summary:**

The invention consists of a web-based service that monitors changes in user-specified web site content. Users are given an account where they specify a list of web pages (Uniform Resource Locators [URLs]) to be monitored with associated keywords of interest for each URL. The user interface with our system is the user's web browser pointing to the URL of our service. The user can select how often each specified URL is checked for changes (sampled) and the method(s) of detected change notification (e-mail, Personal Digital Assistant [PDA], pager, or a near real-time graphical status display). The user can specify the depth of intra-domain hyperlinks that the service will search for occurrence of keywords. The invention uses a web server (Apache) with interfaces to

a database, C programs (Common Gateway Interface [CGI]) and Java programs (servlets with JServ).

### **Invention Description:**

A Service Web Site is provided to the user to allow him to select URL's; corresponding keywords for each URL; the depth to which links will be followed for keyword searching; the frequency of checking for each URL expressed in minutes, hours, or days; the email, pager, and PDA addresses to which notification reports will be sent; the category to which the URL will be assigned; the keyword Boolean expression that will be used to search the web pages. The Boolean expression allows keywords to be joined with AND and OR operators. Once the URL and its parameters are defined, the user then can launch or terminate the search and detection process for each specified URL via the web interface.

The search and detection software is implemented as a Search Daemon, which runs as an independent background process on the host machine. As soon as a daemon is launched, it follows the procedure described below.

### **Data Acquisition and Formatting**

A socket connection is established to the user-specified URL, and a request is sent on the socket to download the HTML from the URL. All the characters sent in response to the request are saved in a file. In addition, a second file is created which contains the formatted version of the text without HTML tags. To create this file, while the characters are being received, any text that is part of an HTML tag is not written to the text-only file. All other characters are written to the file. Thus after all the HTML is received from the URL, the text-only file contains all the text from the URL minus the HTML. During the HTML acquisition, a list of all URL links that appear in the web page is created for later use.

### **Change Detection**

Changes are detected based on a comparison of the previous version of the web page (stored in the database) with the newly downloaded version of the page. The text-only file for a page (described in *Data Acquisition and Formatting* above) is further processed by replacing all white space between words with a single blank character, producing the formatted text version of the page. The new formatted text is then compared to the formatted text of the previous version. If the two do not match then further comparison is required in order to avoid reporting of trivial changes that the user would not be interested in. The keyword counts for the new page are determined. If any one of the keyword counts for the new page differs from the corresponding keyword count for the previous version, then a change is declared between the two versions. After the initial comparison between the previous version in the database and the new version is done, the previous version of the page in the database is replaced by the formatted text of the new version.

### **Pseudocode:**

Produce the formatted text F of the new web page;

Retrieve the formatted text P of the previous version of the web page from the database;

IF (at least one character of F is different from P) THEN

```
{
  Replace P in the database with F;
  Retrieve the boolean keyword expression Exp from the database;
  Search the new page F using Exp;
  IF (Exp is true in F)
  {
    Get the keywords for the URL from the database;
    For each keyword W:
      Count the number of occurrences of the keyword in F;
      Retrieve the keyword counts for the previous version P from the database;
      IF (at least one keyword count for the new version is different from the
        corresponding keyword count for the previous version) THEN
      {
        A change between previous version and new version has been detected;
        Replace the keyword counts of the previous version in the database with
        the
          keyword counts of the new page;
      }
      ELSE
        No change is detected;
    }
    ELSE
      No change is detected;
  }
  ELSE
  {
    No change is detected;
  }
}
```

### Link Traversal

If a change in the web page has been detected and after keyword counts have been tabulated, the software then examines all URL links in the list of links created from the HTML (see *Data Acquisition and Formatting* above).

Pseudocode:

Let LinkList be the list of URL links from the previously searched page, Level = 0, and Limit = the link search depth previously specified by the user.

Begin List Traversal:

```
IF (Level < Limit)
{
  For each link L in the LinkList:
  {
    IF (L is contained in the same domain as the top-level URL) THEN
```

Make a recursive call to the Search Daemon for L:

Download the HTML for L;

Create a new list LinkList2 of all URL links contained in the page for L;

Create a text-only file from the HTML;

Create a formatted text version F of the text-only file;

For each keyword W associated with the original top-level page:

```
{
    Count the number of occurrences N of W in F;
    Add N to the total keyword count T;
}
```

```
}
Store T and L in the database;
```

```
Set LinkList ← LinkList2
```

```
Level = Level + 1;
```

```
Goto: Begin List Traversal;
```

```
}
}
```

### Change and Keyword Hit Notification

When a true change between the new version and previous version is detected, the results are presented to the user in two different formats. First, an electronic message is created and sent to one or more of the user's email address, pager, or PDA (depending on what reporting options were chosen). This message has the following format:

\*\*\*\*\* SMARTCOW ACTIVITY REPORT \*\*\*\*\*

Hey, you got a hit!

URL: <URL name>

KEYWORDS AND HITS:

<keyword 1> : <number of hits 1>

<keyword 2> : <number of hits 2>

<keyword 3> : <number of hits 3>

...  
<keyword N> : <number of hits N>

ABSTRACT:

<abstract A>

<abstract B>

<abstract C>

For more info please log in to SMARTCOW at <URL of service web site>

Happy Milking!!!

All keyword counts are shown, and up to three abstracts from the text are shown. The abstracts are chosen based on the three keywords with the highest frequency of occurrence. If there are at least three keywords, then an abstract is produced for each of

the top three highest frequency keywords showing the first occurrence of each keyword in the text. If there are two keywords, then there will be two abstracts showing the first two occurrences of the most frequently occurring keyword (if it occurs at least twice) and one abstract showing the first occurrence of the other keyword. If there is only one keyword, then there will be up to three abstracts showing the first three occurrences of the keyword, or less if the keyword occurs less than three times. An abstract consists of the ten words preceding the keyword, then the keyword, then the ten words following the keyword.

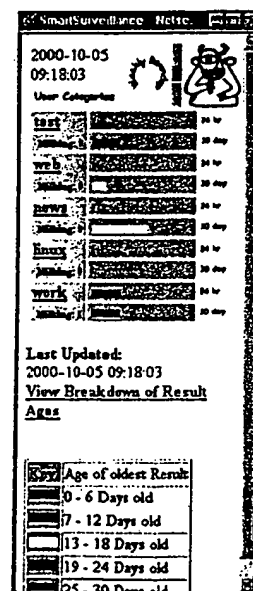
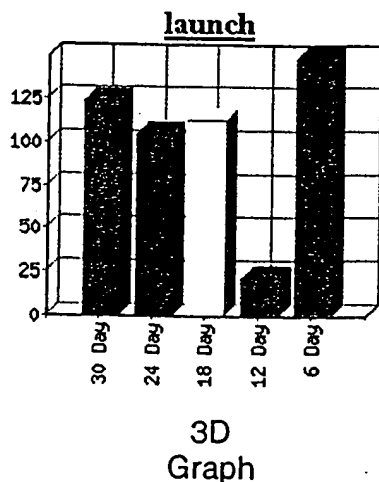
In addition to the electronic notification, the user can view the results at the service web site. An HTML page displaying a format similar to the electronic version is available to the user. Another page is provided to view the total keyword counts obtained from searching URL links that were followed from the top-level or subsequent lower-level pages (see *Link Traversal* above).

### Graphical Displays

The near real-time graphical status display consists of two pop-up windows that show the user two-dimensional (2D) and three-dimensional (3D) graphs the update every 60 seconds.

The 2D graph shows the number of hits per category and the age of the data. The bars of graph are color coded to show aging. The combination of size and color shows the user the activity and the age of the oldest data for that category. A separate black bar for each category shows the 24-hour status. Each bar in the graph is clickable and will bring up a new window showing either the category, 24 hour, or 30-day results depending on which part of the graph is clicked. A single green bar at the top of the window will take the user to the results for all categories.

A 3D display window shows the user the breakdown of hits and separates them into 6-day intervals. This gives the users an instant view showing the number of hits for the past 6, 6 to 12, 12 to 18, 18 to 24, and 24 to 30 days. A separate 3D graph is shown for each category. Each 3-d graph is clickable and will take the user to the category results.



Search Completion

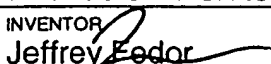

After all processing for a particular top-level URL is completed, including traversal and analysis of all links contained in the top-level and lower-level pages, the Search Daemon then sleeps for a period of time equal to the frequency interval that was specified by the user. If the user has chosen to terminate the processing of the Search Daemon, then the daemon exits at this time.

**Prior Disclosures:** No disclosures have been made outside of Aerospace.

**Exploitation:** Our use of combining URL and keyword used in conjunction with our constrained hyperlink traversing (web crawling) and our change detection algorithm offers a unique and effective method of discovering changes in information pertinent to a user's specified interests.

**Funding:** JO 8511-17

PLEASE WRITE SIGNATURES AND DATES IN BLACK INK

INVENTOR Andrew Quintero	DATE 10-5-2000	INVENTOR 	DATE 10-5-2000	INVENTOR 	DATE 10-5-2000
ORGANIZATION AND CCC EPD - 6807		ORGANIZATION AND CCC CSD - 5814		ORGANIZATION AND CCC CSD - 5814	
INVENTOR Karen Richardson	DATE 10-5-2000	INVENTOR Donald Scott	DATE 10-5-2000	INVENTOR Ken Piper	DATE 10-5-2000
ORGANIZATION AND CCC SD - 5814		ORGANIZATION AND CCC CSD - 5814		ORGANIZATION AND CCC CSD - 5814	
WITNESSED, READ AND UNDERSTOOD BY:		DATE:	AND BY:	DATE:	

OCT 5 - 2000